Donnees et creation de valeur

TL;DR

- **La donnée est le seul différenciateur de l'IA :** l'algorithme et la puissance de calcul sont devenus des commodités. La donnée, elle, est unique et impossible à racheter.
- **Exploitée stratégiquement, elle crée des barrières inacceptables :** data moat, flywheel. Ces mécanismes se construisent - ils ne s'achètent pas.
- **Mal gouvernée, elle crée du risque :** biais, fuite, non-conformité, sanctions. C'est la responsabilité du manager - pas du DSI.

1) La donnée, matière première de l'IA

Les 3 ingrédients de l'IA - et leur statut stratégique

Ingrédient	Statut aujourd'hui
Algorithme	Commodité - accessible via API (Open AI, Mistral, Google…) pour quelques centimes
Puissance de calcul	Commodité - louable a la demande (AWS, Azure, OVHcloud, Scaleway)
Données	Différenciateur unique - propriétaires, non reproductibles, impossibles à acheter

Principe clé : quand quelque chose devient une commodité, l'avantage se déplace vers ce qui reste rare. La fenêtre pour construire un avantage data est ouverte - elle ne le sera pas indéfiniment.

Garbage in, garbage out

Avoir de la donnée ne suffit pas. Une donnée biaisée ou incorrecte produit des décisions biaisées, automatisées à grande échelle. La qualité de la donnée est une responsabilité managériale, pas technique.

Exemple : un modèle de scoring entraîné sur des données historiques biaisées, reproduit et amplifie ces biais sur des milliers de décisions par jour.

La chaine de valeur de la donnee

Etape	Ce qui se passe
1. Collecte	Captation à la source : formulaires, capteurs, transactions, logs
2. Stockage	Entrepôt de données : data warehouse, data lake, bases départementales
3. Nettoyage	Suppression des doublons, correction des erreurs, uniformisation
4. Analyse	Exploration, recherche de patterns, formulation d'hypothèses
5. Modèle IA	Construction d'un système prédictif ou de recommandation
6. Décision	Recommandation qui alimente une action concrète
7. Valeur	Résultat : efficience, avantage compétitif, nouveau service

> **Chiffre clé**
> Les data scientists passent 60 à 80 % de leur temps à préparer les données - pas à modéliser. Le problème n'est pas l'algorithme : il est dans l'organisation de la donnée en amont.
>
> NB : ce chiffre tend à être GRANDEMENT diminué avec l'utilisation de LLMs (Ex : Claude Code intégré à des outils tel que Databricks)

Le problème des silos

Marketing, Finance, Operations, RH - chaque département détient ses données dans ses propres systèmes, avec ses propres définitions. Ces systèmes ne se parlent pas.

Conséquence : impossible de croiser les sources pour un projet IA. La solution n'est pas technologique - c'est une décision organisationnelle. Acheter une plateforme d'intégration sans gouvernance préalable créé un silo centralisé, pas une solution.

2) L'avantage compétitif par la donnée

Le data moat - le fossé défensif

Concept issu de la strategie competitive (Warren Buffett) : un “moat” protège durablement un avantage concurrentiel. La donnée est le moat le plus puissant car impossible à acheter rapidement. On peut copier un algorithme - on ne peut pas racheter 5 ans de données comportementales.

Entreprise	Data moat
Amazon	25 ans de comportements d'achat de 300 M de clients - granularité inégalée
Doctolib	Part majeure des rendez-vous médicaux en France - vision unique
Tesla	Milliards de km de conduite reelle remontés en temps réel par des millions de véhicules

Pattern commun : l'avantage n'est pas dans la technologie. Il est dans la donnée accumulée sur le long terme, structurellement impossible à répliquer rapidement.

Le flywheel - la boucle vertueuse

> **Mécanisme**
> Plus d'utilisateurs -> plus de données -> meilleur modèle IA -> meilleur produit -> plus d'utilisateurs.
>
> La roue tourne et s'accélère. Impossible à stopper - et à répliquer pour un entrant.

- **Netflix** — Plus tu regardes → meilleures recommandations → tu regardes plus → plus de données sur ce qui retient → meilleures recommandations → ♾️
- **Spotify** — Plus tu écoutes → Discover Weekly s'affine → tu découvres des titres que tu aurais cherché ailleurs → tu restes sur Spotify → plus tu écoutes → ♾️
- **TikTok** — Chaque seconde de vidéo regardée ajuste le fil en temps réel → vidéo suivante plus pertinente → tu restes → plus de données comportementales → ajustement encore plus précis → ♾️
- **BlaBlaCar** — Plus de trajets → plus de notations → confiance accrue → nouveaux utilisateurs attirés → plus de trajets → ♾️

Les 4 modèles de création de valeur

01	**Monetisation directe** Vendre la donnée agrégée à d'autres organisations. Ex : Nielsen, Dun & Bradstreet.
02	**Amélioration produit** Utiliser la donnée pour personnaliser et fidéliser. Ex : Netflix, Amazon, Spotify.
03	**Efficience opérationnelle** Optimiser ses propres opérations. Ex : Michelin (maintenance prédictive), SNCF (voies ferrées).
04	**Nouveau service** La donnée crée une offre inexistante. Ex : John Deere (intelligence agronomique).

La grille Valeur x Risque x Dépendance appliquée

Dimension	Ce qu'elle révèle
Valeur	Avantage compétitif durable, barrière à l'entrée croissante, accélération permanente
Risque	Vol (cybermenace), biais (décisions injustes), non-conformité (RGPD, AI Act)
Dépendance	Hebergeur etranger (Cloud Act), captivité technologique, perte de contrôle sur son moat

Un manager qui optimise une dimension sans tenir compte des deux autres prend une décision incomplète. Toujours les trois - simultanément.

3) Gouvernance & qualite des donnees

Ce que gouverner la donnée signifie - 3 questions

1	**Qui a le droit de toucher à quelle donnée ?** Lecture, modification, suppression, partage externe - droits d'accès et responsabilité.
2	**Dans quel état cette donnée doit-elle être ?** Définition de la qualité, règles de validation, processus de correction.
3	**Qui est responsable quand ce n'est pas le cas ?** Responsabilité, comptes à rendre, conséquences.

La gouvernance n'est pas un sujet technique. C'est un sujet de pouvoir. Cela ne se résout pas avec un meilleur outil - cela se résout avec une décision managériale claire.

Les 4 problèmes chroniques

01	**Silos** Chaque département a ses données avec ses définitions. Impossible de croiser sans semaines de réconciliation.
02	**Qualite** Doublons, valeurs manquantes, formats inconsistants. Anodin individuellement - catastrophique sur un modèle prédictif.
03	**Tracabilite (data lineage)** D'où vient cette donnée ? Qui l'a modifiée ? Sans réponse, impossible de valider ce qu'un modèle apprend.
04	**Conformité** RGPD (4 % du CA mondial), AI Act, réglementations sectorielles. Un usage non autorisé expose l'organisation et le décideur.

Les 2 rôles à connaître

Rôle	Ce que c'est
Data Owner	Responsable métier d'un périmètre de données. Décide quoi collecter et dans quel but. Ce rôle est le votre - quel que soit votre intitule.
Chief Data Officer	Pilote la stratégie données à l'échelle de l'organisation. Arbitre entre les métiers. Rôle stratégique - pas technique.

Point commun : ce sont des décideurs, pas des exécutants.

Le data catalog

Un annuaire interne pour les données : contenu, localisation, Data Owner, droits d'accès, cadre légal. Sans catalogue : chaque projet IA démarre en mode archéologie (semaines perdues). Avec catalogue maintenu : quelques heures.

Les 3 questions du manager avant tout projet IA

Quel que soit l'outil, le prestataire ou le budget - posez ces trois questions avant de lancer n'importe quel projet IA. Et obtenez une réponse claire à chacune.

01	**La donnée existe-t-elle ?** En quantité et qualité suffisantes pour le cas d'usage vise - pas 'pourrait-on la reconstituer'.
02	**Est-elle propre et exploitable ?** Ou nécessite-t-elle 6 mois de préparation ? Si oui, c'est ce qu'il faut planifier et budgéter.
03	**A-t-on le droit de l'utiliser ?** Consentement, contrats fournisseurs, réglementations sectorielles - pour ce cas d'usage précis.

A retenir

Si la réponse à l'une de ces trois questions n'est pas 'oui' clairement - le projet n'est pas prêt. Ces questions se révèlent à chaque étape significative, pas seulement au lancement.